The use of the Fifteen Factor Questionnaire as a Pre-selection Tool: Differential Item Functioning Based on Gender and Ethnicity

---

A thesis submitted in partial fulfilment of the requirements for the

Degree of

Masters of Science in Applied Psychology

in the

University of Canterbury

by

Nicholas John Catto

University of Canterbury

2008

Table of Contents

List of Tables and Figures

Acknowledgements

The author wishes to express his sincere appreciation to Doctor Linda Trenberth and Miss Teresa Macgregor for their assistance and support in the preparation of this manuscript and for kindly acting as supervisor and co-supervisor.

Abstract

The use of psychometric testing during pre-employment selection processes is increasing within New Zealand and around the world. Research into potential measurement bias amongst these tests has lagged behind that of other tests such as measures of cognitive ability. This study explored item level measurement bias through differential item functioning (DIF) analysis. Differential item functioning analysis was conducted based on gender and ethnicity in a pre-employment selection tool, the Fifteen Factor Questionnaire Revised Edition (15FQ+; Psytech, 2002). The current study used a sample of 4798 participants who had completed the 15FQ+ and voluntarily identified themselves as male or female and New Zealand European or Maori. An item response theory (ITR) framework was used to study DIF which was assessed using ordinal logistic regression. This study found that 62.35% of items displayed some degree of DIF based on gender and 40.59% of items displayed some degree of DIF based on ethnicity. The current study also found that 20% of items displayed no DIF, whilst 23.53% of items displayed DIF across both gender and ethnicity. This is similar to other research based on DIF across a number of domains where psychometric testing is used.

1 Introduction

An exploration of psychometric testing in employment and an investigation into differential item functioning in a pre-employment selection tool

Psychometric tests are used with individuals on a regular basis to make important decisions in a diverse range of settings, from employee selection, placement, and development through to patient and client assessment and treatment. Psychometric testing occurs in a wide range of environments, from schooling and education, through to employment and imprisonment. A whole host of human behaviours, traits, abilities, tendencies, knowledge and psychopathologies can now be psychometrically tested. The measurement of an individual's knowledge, abilities, attitudes, behaviours, mental status and personality traits has now become widespread around the globe (Mitchell, 1999). Psychometric testing is not limited to the Western World, with psychometric testing increasing in countries as diverse as China and India (Tyler & Newcombe, 2006). It remains unlikely for an individual to go throughout their life without having some experience with psychometric testing (Roznowski & Reith, 1999). The purpose of this current study is to investigate differential item functioning (DIF) in a pre-employment selection tool, using ordinal logistic regression (OLR; Roznowski & Reith, 1999; Zumbo, 1999). This study will highlight the increasing importance personality plays in employment, and why its popularity has increased globally. This study will also highlight that measurement bias has, to date, been limited in its exploration of measurement bias in pre-employment selection tools. Finally this study will demonstrate one method of detecting measurement bias, and highlight its importance and relevance.

The use of psychometric tests as a form of pre-employment assessment to aid in finding the most suitable person as a potential employee has increased world wide in the past two decades (Jenkins, 2001). Psychometric testing increasingly occurs during pre-employment selection and assessment of potential employees (Jenkins, 2001). One of the most common psychometric tests in the pre-employment arsenal is the use of a personality inventory. Psychometric testing which asses an applicant's personality now forms and informs many pre-employment selection processes (Jenkins, 2001).

Research supports the limited utility of personality assessment in the selection of potential employees (Tyler & Newcombe, 2006). Psychometric testing, though having been around for over 100 years has only recently increased in popularity and become widespread in New Zealand (Sheppard, Han, Colarelli, Dai, & King, 2006; Ones & Anderson, 2002; Jenkins, 2001). Across a broad spectrum of jobs both skilled, semi-skilled and unskilled and organisations from corporate and government through to non-governmental organisations there has been a substantial increase in the use of pre-employment personality testing as a means of hiring potential employees (Ones & Anderson, 2002). Further to this, it has become big business, as pre-employment selection tools such as the Occupational Personality Questionnaire (Bartram, Brown, Fleck, Inceoglu & Ward, 2006) and the Hogan Personality Inventory (Hogan & Hogan, 1992) generate millions of dollars in revenue each year for the consultancy companies that administer and distribute them and the organisations that create, develop and publish them (Ones and Anderson, 2002).

Jenkins (2001) suggests three reasons why psychometric testing is more than just a managerial fad and will continue to grow. Firstly, the costs associated with engaging a consulting

organisation specialising in psychometric assessment represent considerable business expenditure. Further to this there are the significant costs and financial commitments associated with training employees to use and administer tests. Secondly any licensing fees and ongoing costs associated with becoming accredited to use a publisher's tool also require an on going financial commitment from the organisation. Thirdly, a genuine belief exists among organisations in their suitability and ability at detecting job applicants with the correct skills set and key attributes. Indeed, "...HR professionals [are provided] with a myriad of choice on what psychometric assessment to use...in what has now become a very commercially aggressive industry..." (Englert, 2006). For example the New Zealand Council for Educational Research (NZCER, 2008) product manual for human resource assessment contains over 50 different tests (not including different formats, versions and subtests) that measure the whole gamete of an individual's abilities, traits, knowledge, and behaviours that may be required for the work place. Jenkins (2001) suggests that these factors will ensure that psychometric testing in employment will be cemented in organisations for the foreseeable future.

1.1 Psychometrics and Job Performance

Research shows that some personality dimensions such as conscientiousness and honesty are valid predictors of job outcomes in certain circumstances (Barrick & Mount, 1991; Tett, Jackson & Rothstein, 1991). Timmerman (2004) found correlations between NEO PI-R (Costa & McCrae, 1992) Conscientiousness (r = .16) and Agreeableness (r = .16) and supervisory performance ratings. Irrespective of their correlations with work place performance which are well known, characteristics such as honestly, dependability, conscientiousness and agreeableness

are highly valued by prospective employers (Tyler & Newcombe, 2006; Ones & Anderson, 2002). Pre-employment personality inventories represent one of the most effective ways to objectively measure and asses these characteristics in potential employees (Ones & Anderson, 2002). Jenkins (2001) identifies three reasons for the organisational rationale behind psychometric testing. Firstly, psychometric testing is used in employment by organisations because of their perceived objectivity. Secondly, their predictive abilities are now widely reported and publicised extensively. Thirdly a belief exists among organisations in their ability to siphon off unsuitable job applicants. In conjunction with other assessment such as cognitive ability testing, structured behavioural based interview questions and integrity testing the incremental validity and predictive power is stronger than when any one assessment is used in isolation (Sheppard, Han, Coralli, Dai & King, 2006; Barrick & Mount, 1991; Tett, Jackson & Rothstein, 1991; Day & Silverman, 1989). This large body of evidence suggests that quality personality inventories appear to extract a degree of variance in work place performance not capable of being measured by any other current human resource tool (Tyler & Newcombe, 2006; Jenkins, 2001).

Though personality as a means for making employment decisions has not been without controversy or contention personality measurement is now widely used in pre-employment selection (Jenkins, 2001). It seems likely that personality assessment based on a Five-Factor Model adds incremental validity in workplace selection that cannot be accounted for by any other human resource tool or method (Tyler & Newcombe, 2006). The continued use of psychometric testing in employment likely reflects the view that testing in general and testing of personality will continue to play a significant part in the prediction of work place behaviour and

performance (Jenkins, 2001; Tett, Jackson, & Rothstein. 1991; Schmidt & Hunter, 1998). Sheppard et al. (2006) suggests that as their popularity increases so their gatekeeper function will also increase in employment environments. Given their increasing importance and the continuing reliance upon personality measures in employee selection and the danger that they can become a means solely by which an individual is hired or rejected a number of concerns have been raised (Escorial & Navas, 2007). In particular, there is increasing concerns about potential measurement and test biases that may exist (Sheppard et al 2006; Sackett & Wilk, 1994).

1.2 Measurement Bias and the 'Appropriate Taxonomy'

The decisions based on psychometric tests have a significant personal, social and political impact on the individual and society as a whole (Clauser & Mazor, 1998). Despite testing being developed and "...intended to lead to objective decisions about individuals, they have been found to have an adverse impact on different groups..." (Roznowski & Reith, 1999 p. 248). Test bias can trace its history back to the Han Dynasty (202 B.C. – 220 A.D.) when scribes rewrote examinees written answers before being graded and thereby ensuring anonymity and eliminating but one potential form of bias (Holland & Wainer, 1993). The differential performance of individuals based on race, gender, ethnicity, socioeconomic status, disability and other extraneous factors is now widely researched and the claims of discrimination and bias are now widely publicised (Roznowski & Reith, 1999).

Personality inventories and tests have mainly been used in clinical settings and have only recently begun to have an impact on personnel selection (Sheppard et al. 2006). Most studies

focus on the bias of personality measures that assess psychopathology such as the MMPI or anger such as the Aggression Questionnaire (Sheppard et al. 2006; Condon, Morales-Vives, Ferrando & Vigil-Colet, 2006). Whilst there has been a long-standing concern among researchers and practitioners about differentiated predictions in employment oriented personality measures based on ethnicity and gender, the issue has remained largely unexplored (Saad & Sackett, 2002; Sackett & Wilk, 1994). Further to this, a lack of consensus over what constitutes personality and an appropriate taxonomy of personality traits has lead to less research in personality tests and even less on pre-employment personality measures and few studies examining non-clinical populations (Sheppard et al. 2006; Ones & Anderson, 2002; Goldberg, 1990).

Sheppard et al. (2006) suggests this lack of research is a result of the historical difference in the use of personality tests and ability tests. Ability tests and IQ tests operate as the gatekeeper function to higher education and some employment opportunities. Ethnic differences in scores on ability tests have resulted in different admission rates to higher education and hiring rates. This has lead to an increase in research and policy to examine group differences and asses whether these differences are due to test biases or true differences. Ones and Anderson (2002) go so far as to say that researching the differences in cognitive ability tests has all but consumed researchers at the expense of personality measures. In 1994 Sackett and Wilk (1994) were unable to locate any studies which addressed the issue of differential prediction using personality measures in the employment environment. Borsboom, Mellenberg and van Heerden (2002) indicate that the detection of test bias is of equal importance in the field of personality psychology as any other domain of psychological measurement. In light of the increasing use of pre-employment

personality inventories used during selection in making high stakes decision this statement seems increasingly important given the current high use environment.

The development of psychometrically sound personality inventories for non-psychiatric populations such as the OPQ (Bartram et al. 2006) and HPI (Hogan and Hogan, 1992) and the emergence of the Big Five and other taxonomies have resulted in an increased consensus emerging as to what constitutes a normal personality structure (Sheppard et al. 2006; Goldberg, 1990; McCrae & Costa, 1985; Norman, 1963; Catell, 1946). For example, McCrae and Costa (1997) found a Five-Factor Model displayed a similar structure across German, Portuguese, Hebrew, Chinese, Korean and Japanese translation of the NEO-PI (Costa & McCrae, 1992) whilst Collins and Gleaves (1998) established a moderate fit for a Five-Factor Model for African American and Caucasian job applicants. This unifying ground, despite its short comings has provided theorists and practitioners a base from which to study, communicate and utilise personality in the workplace (Tyler & Newcombe, 2006). It is now possible to assess potential bias in personality inventories to the same degree that has been applied to cognitive ability measures (Sheppard et al. 2006). This has allowed an increase in psychometric rigor to be applied to personality inventories, which matches the rigor applied to cognitive ability measures (Sheppard et al. 2006). In determining the accuracy of scales such as pre-employment personality measures understanding how different groups perform on a scale is extremely important (Collins, Raju & Edwards, 2000).

Test validity remains central to test theory and scientific progress, but there are also a number of ethical, legal and political issues that colour the issue of test use (Borsboom, Mellenberg & van

Heerden 2002). That a test is not biased is an important consideration in the selection and use of any psychological test. Holland and Wainer (1993) identify three central themes in ensuring test fairness and effectiveness. Firstly the review the items receive from subject matter experts from major sub groupings such as gender and ethnicity during test development. Additionally the comparisons of predictive ability based on subgroup membership. Finally statistical analysis of the performance of subgroups relative to each other, at the over all test level, as well as the item level analysis. Test measurement, however, is never perfect and errors will always arise in testing (Zumbo, 1999). Sheppard, Han, Coralli, Dai and King (2006) identify test bias as "...psychometric inequalities among sub-groups [that] can take the form of relationship bias or measurement bias..." (p. 443). Relationship bias is the association between a test score and an external criterion measure. This occurs when an individual from a subgroup has an equal test score but an unequal probability of success on the criterion. For example, an individual may perform capably on the criterion measure but have a lower probability of having a passing score on the predictor. Measurement bias is the property of the test item itself. A test is biased if an individuals has the same latent trait but an unequal opportunity of the same test score (Sheppard et al. 2006). Biases such as these can lead to systematic errors that distort and erode inferences made in employment selection (Ones and Anderson, 2002; Zumbo, 1999).

The items that comprise an overall test or scale for assessing a psychological construct should depend on the participant's level on the variable being measured and not on irrelevant characteristics (Fidalgo, Hashimoto, Batram & Muniz, 2007). A test must be fair to all applicants and should not be biased toward a subgroup of the application pool or population (Zumbo, 1999). Employment processes within New Zealand and around the world are governed by a

number of laws and guidelines which are designed to minimise disparities that could potential occur in the workforce based on gender and ethnicity. For example, the New Zealand Human Rights Act 1993 requires that an employer demonstrate that their selection practises (not limited to psychometric testing) do not discriminate based on factors such as age, gender and ethnicity. Equal Opportunities in Employment requires that selection methods offer equal employment opportunities to all job applicants and the onus remains on the employer to show that selection methods are both fair and job related (Ones & Anderson, 2002).

What though of any potential biases that may arise because of using pre-employment personality measures? Collins, Raju, and Edwards (2000) identify that measuring and understanding of how different groups respond is one of the most important areas in determining test accuracy. Though the biases in cognitive ability testing remain widely known and extensively research this issue remains largely unexplored in personality testing (Sheppard et al. 2006). The consequences of adverse impact are especially serious in countries such as the United States and Britain such that if it can be shown to be discriminatory in practice an organisation may be liable for compensatory awards (Ones & Anderson, 2002). Along with these pecuniary considerations, there are also a number of fiduciary and ethical considerations which must be factored in when a practitioner is considering test usage.

1.3 Differential Item Functioning

One way to test measurement bias and item bias is through differential item function (DIF; Sheppard et al. 2006; Zumbo 1999; Clauser & Mazor, 1998). This involves examining

differences in tests at an item level as opposed to the overall test level (Roznowski & Reith, 1999). Differential item functioning (DIF) is a common method used to evaluate invariance at the item level and for assessing whether an item reflects the construct of interest and not measurement irrelevancies, such as those based on gender and ethnicity (French & Maller, 2007). Differential item functioning is often used to examine group differences based on ethnicity and gender. Differential item functioning can be used for a range of different group comparisons such as tests which have been translated in to other languages or with individuals with another difference such as a disability or cultural and ethnic background (Abedi, Leon & Kao, 2007).

Differential item functioning occurs when individuals from different subgroups, such as gender or ethnicity though having the same amount of latent trait, such as cognitive ability or extraversion, have a different probability of giving a certain response to an individual item on a test. Differential item functioning is displayed when individuals from different subgroups with the same underlying true ability or trait have a different probability of giving a certain response to an individual item (Sheppard et al. 2006; Borsboom, Mellenbergh & van Heerden, 2002; Collins, Raju & Edwards, 2000; Zumbo, 1999). For example two individuals may have equal general mental ability overall when measured, and show no difference in the probability of correctly answering an individual item that makes up that test, no DIF is present. When individuals based on group membership, such as ethnicity or gender have the same level of a latent trait have a differing probability of responding correctly or endorsing that individual item measuring that trait, the item is said to display DIF. The item is differential based on an irrelevancy such as a gender, or ethnicity rather than displaying a uniform response pattern. That

item may subsequently advantage the group which displays the higher probability of endorsing or correctly answering that item (Scherbaum & Goldstein, 2008).

The detection of DIF indicates that an item may not be performing uniformly across groups (Collins, Raju & Edwards, 2000; Zumbo, 1999). Therefore the item has lower construct validity for one group, because it is tapping an extraneous factor in that group (Holmes-Finch & French, 2008). The assumption of DIF is that it has a detrimental effect on the meaning of test scores and on the measurement of the underlying trait for that subgroup (Roznowski & Reith, 1999). This is important because decisions about an individual are based on overall test levels. However, that overall test level result is based on individual test items that comprise that scale (Roznowski & Reith, 1999). If items operate in a differential manner then this would suggest that scores between-group may not be comparable in nature, and could lead to inequitable treatment (Holland & Wainer, 1993). It is important to note that even though DIF may be present, that alone is not enough to reject an item or a test. Indeed, if an item is biased or offensive or irrelevant to both groups, then DIF techniques will not detect bias. Differential item functioning techniques are reliant on detect between-group differences based on differential response patterns (Zumbo, 1999). Instead, Holland and Wainer (1993) suggest DIF techniques should, at the very least guide the development and construction of items and tests.

1.4 Detecting Differential Item Functioning

There are a number of techniques for detecting differential item functioning (Holmes-Finch & French, 2008). Simultaneous Item Bias Test (SIBTEST) and Item Response Theory (IRT)

likelihood ratio tests specifically target DIF. Other general techniques such as logistic regression and the Generalised Mantel-Haenszel test are statistical techniques which can be used to detect differential item functioning (Holmes-Finch & French, 2008). A major determining factor in which method to use is whether the data is binary in nature or ordinal in nature. Binary data has traditionally dominated DIF analysis due to the predominance of studies on general mental ability tests, which are frequently binary in nature (Zumbo, 1999). Differential item functioning determines that the 'reference group' is the group with which a 'focal group' is compared against, the 'focal group' is identified as the group expected or thought to be disadvantaged, and will therefore show DIF.

The current study will use ordinal logistic regression (OLR) which will provide a probability of endorsing a response to an item (Swaminathan & Rogers, 1990; Zumbo, 1999). Ordinal logistic regression is a cost and time effective method of detecting DIF, therefore its use in test construction, development and evaluation is strongly encouraged by its proponents (Zumbo, 1999). Ordinal logistic regression uses an IRT framework to assess between-group differences (Clauser & Mazor, 1998). Differences between the item parameters for two groups are compared. Item characteristic curves (ICC) provide an insight in to how groups respond on individual items. Figure 1 replicates an ICC with no incidence of DIF. Figure 2 demonstrates how a reference group and focus group ICC differ on an individual item and would subsequently be identified as displaying DIF (Clauser & Mazor, 1998). It shows how the threshold for endorsement in a personality measures or responding correctly in an aptitude test varies based on group membership.

Figure 1

Smith & Reise (1998) ICC for a focal and reference group for an item that displays no DIF

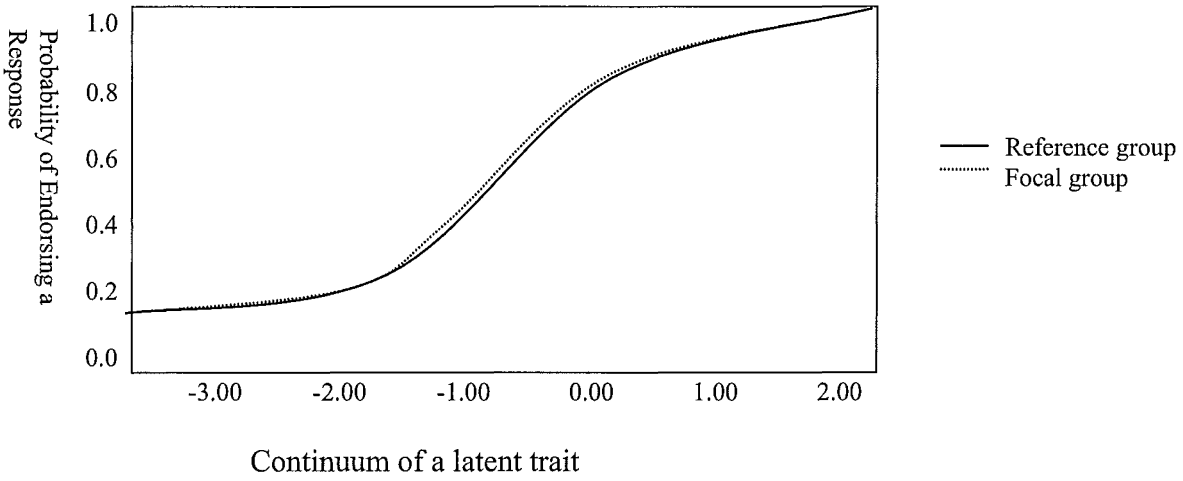

Figure 1 shows that the item characteristics curve for the reference group, and the focal group match or overlap each other. Therefore there is no difference in the threshold for endorsement, or correctly answering an item. The probability of endorsing a certain response remains equal across two subgroups for that individual item.

Figure 2

Smith & Reise (1998) ICC for a focal and reference group for an item that displays DIF

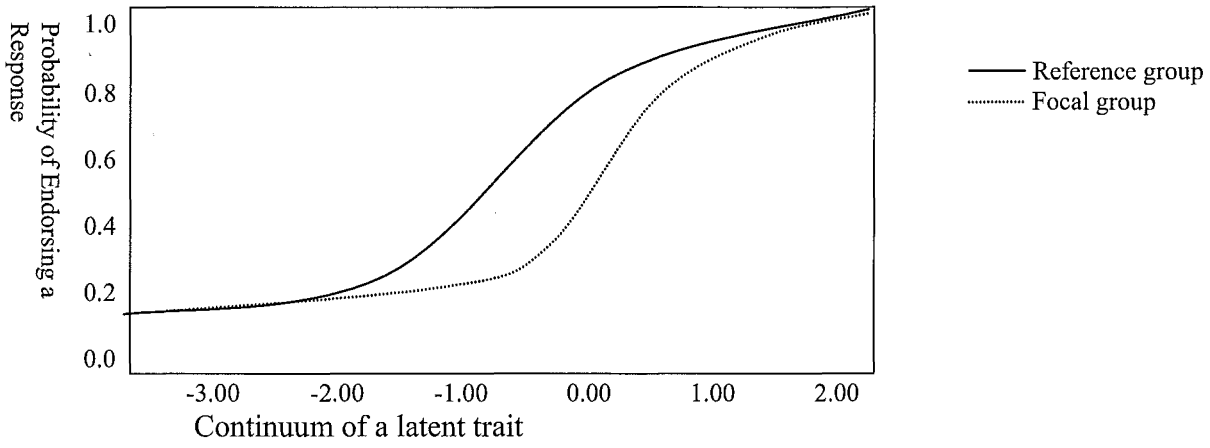

Figure 2 shows that the focal group and reference group response pattern do not overlap, and therefore the item parameters vary based on group membership. The item under examination would be identified as containing DIF because factors such as gender or ethnicity in spite of equal trait levels do not display equal probabilities of endorsing the item under consideration (Smith & Reise, 1998). An individual in the focal group would need more of a latent variable to endorse the item or answer it correctly in the case of an aptitude test, than the individual in the reference group. The item would be consider more difficult in an aptitude measure, or require more of a latent trait in a personality measure for the focal group.

1.5 The current study

Personality testing in pre-employment employee selection is likely to continue, expand and increase in its impact on personnel selection (Sheppard et al. 2006; Jenkins, 2001). As this expansion increases questions will continue to emerge surrounding measurement bias in pre-employment personality inventories (Roznowski & Reith, 1999). The possibility of test bias in employment-oriented personality inventories remains largely unexplored to date (Sheppard et al. 2006; Sackett & Wilk, 1994). Although item bias on ability and achievement measures have received considerable attention, very little work has focused on the detection of item level bias in pre-employment personality measures (Sheppard et al. 2006; Sackett & Wilk, 1994; Thissen, Steinberg, & Gerrard, 1986). One of the most important questions surrounding the use of personality measures used in pre-employment testing environments is whether gender and ethnic groups exhibit differing responses to individual items despite the same level of a latent trait (Ones and Anderson, 2002; Escorial & Navas, 2007). In addition, as the public become more

aware of psychometric testing through personal experience, the popular media and the growing professionalization of the human resources sector questions surrounding their validity, reliability and utility will increase. This environment provides the rationale for the current study that is to be conducted.

1.6 The 15FQ+

The current study will use the Fifteen Factor Questionnaire Updated Version Form A (15FQ+; Psytech, 2002). The 15FQ+ is a revised and updated version of the Fifteen Factor Questionnaire (15FQ; Psytech, 2002) and is designed to provide a comprehensive assessment of a normal personality structure. It is designed for use in the international business environment for the purposes of selection and assessment during pre-employment selection programmes. The 15FQ+ is extensively licensed internationally with a significant market share. The 15FQ+ measures 15 core personality factors that were first identified by Cattell (1946) with the addition of a scale which measures Intellectance. The 16 Primary Factors scales are designed to measure the following core personality structures, Expedient/Conscientious, Hard-Headed/Tender-Minded, Concrete/Abstract, Conventional/Radical, Informal/Self-Disciplined, Affected by feelings/Emotionally stable, Self-Assured/Apprehensive, Composed/Tense-Driven, Distant Aloof/Empathic, Retiring/Socially Bold, Group-Oriented/Self-Sufficient, Low Intellectance/High Intellectance, Accommodating/Dominant, Direct/Restrained, Trusting/Suspicious, and Sober Serious/Enthusiastic. The 15FQ+ consists of five Global Factors, Extraversion, Anxiety, Pragmatism, Independence and Self Control which are similar to 'The Big Five' (Tyler & Newcombe, 2006; Psytech, 2002). The 16 Primary Factors scales positively or negatively load

on to the five Global Factors, and provide more general insights in to an applicant's personality. Further to this, the 15FQ+ has additional scales built in to measure Emotional Intelligence, Work Attitudes and Social Desirability, Central Tendency, and Infrequency. The 15FQ+ can be completed in Short Form (100 questions) or Long Form (200 questions).

The purpose of this study is to examine the extent of potential measurement bias based on gender and ethnic subgroups in an inventory that asses dimensions of normal personality function and is used as a pre-employment selection tool. This represents an area that has been under studied to date at an international and national level (Sheppard et al. 2006). Further to this, the pre-employment selection tool under study has not been subjected to a DIF analysis. The objectives of this study are to identify the number of items that show gender or ethnicity DIF. Secondly, this study aims to asses the differences in endorsement of items by the focal group by establishing which items are more likely or less likely to be endorsed by the focal group when compared with the reference group and matched for overall trait level. Thirdly, the number of items that display no DIF will be determined. Finally the number of items that display across subgroup DIF will be determined.

2 Method

2.1 Participants

A private consulting organisation using the 15FQ+ (Psytech, 2002) as part of a personnel selection programme provided the data for this study. Participants were selected for this study based on whether they voluntarily indicated their gender as male or female, and as either New Zealand European or Maori at the time of completing the 15FQ+ (Psytech, 2002). As participation was part of a selection programme when applying for a job, no compensation was provided. Informed consent was gained before the commencement of the 15FQ+ (Psytech, 2002; Appendix I), including the provision that data would be retained on a database and could be made available for future research (Appendix I). A total of 4798 participants (mean age = 33.4, SD = 9.8, range = 16-68) who voluntarily identified themselves as Maori or New Zealand European and male or female were extracted from a larger database of individuals who had completed the 15FQ+ (Psytech, 2002). There were 2919 male participants (mean age = 33.6, SD = 9.9, range = 17-68) and 1879 female participants (mean age = 33.0, SD = 9.5, range = 16-65). Of the 4798 participants 579 voluntarily identified themselves as Maori (mean age = 32.5, SD = 9.4, range = 17-61) and 4219 identified themselves as New Zealand European (mean age = 33.5, SD = 9.9, range = 16-68). Participants came from a wide range of skilled, semi-skilled, and unskilled occupations, and a wide variety of employment environments such as government organisations, not for profit organisations, corporate and private organisations.

2.2 Materials

The 15FQ+ Form A (Psytech, 2002) consists of 16 Primary Factors scales, made up of 200 items. Each of the 16 Primary Factors scales has twelve questions. One hundred and eighty three questions are keyed "True (A), Uncertain (B) or False (C)" or "Often (A), Sometimes (B), or Rarely (C)" (1; 2; 3;). For example, "I have never broken anyone's trust or confidence". The remaining 17 questions are keyed between a choice of two activities preferred or uncertain. For example "I more admire the work of: famous engineers and scientists (A), uncertain (B), great artists and philosophers (C)" (1; 2; 3). Each of the 16 Primary Factor scales loads positively or negatively on to Global Factors scales, Extraversion, Anxiety, Pragmatism, Independence and Self Control. The 15FQ+ (Psytech, 2002) has inbuilt scales which measure Social Desirability, Faking Good and Faking Bad used to determine a participants level of impression management. Additionally, the 15FQ+ (Psytech, 2002) has an Infrequency and Central Tendency scale which measures response style. Two inbuilt scales measure Emotional Intelligence and Work Attitude. The 15FQ+ (Psytech, 2002) 16 Primary Factors scales all have reliability coefficient alpha's above .70 (range .74-.85) indicating high reliability (Psytech, 2002). The test-retest reliability coefficients for the 15FQ+ (Psytech, 2002) 16 Primary Factors scales range from .77 to .89. The 15FQ+ shows convergent validity with the 16PF (Catell, 1946), the NEO PI-R (Costa & McCrae, 1992), and the OPQ32i (Bartram et al. 2006) ranging from r = -.71 to r = .84. The 15FQ+ (Psytech, 2002) 16 Primary Factors scales also shows meaningful correlations with job performance, managerial performance, and absenteeism ranging from r = -.45 to r = .37 (Psytech, 2002).

2.3 Procedure

Participants were read a detailed set of standardised instructions on how to complete the 15FQ+ and given the opportunity to complete a practice example before beginning. Applicants who completed the 15FQ+ were asked to voluntarily and anonymously provide biometric data including their age, gender and ethnicity information at the time of completing the test by marking the appropriate box. Answers are completed by pencil and paper, via computer, or on-line. The test is designed to take thirty minutes, with the opportunity to receive verbal feedback given when the test has been scored.

This study is a descriptive research design and is a between-groups study. The variables have been measured and made available for an empirical investigation. The dependent variable in this study is item response (1, True; 2, Uncertain; 3, False). The grouping variable in this study will be male and female or Maori and New Zealand European. The independent variable in this study will be the total score for each of the 16 Primary Factor scales.

As the 15FQ+ is a commercially available test, the structural matrix and assignment of items to their respective 16 Primary Factors scales was not known to the researchers. Therefore test items had to be assigned to the relevant 15FQ+ 16 Primary Factors scales (Appendix II). Items were assigned to the relevant 15FQ+ 16 Primary Factors using (a) information contained in the 15FQ+ Fifteen Factor Questionnaire Technical Manual (Psytech, 2002); (b) intuitive judgement based on face validity, (c) factor analysis (d) item scale statistics. A total of 170 out of a possible 200 items were assigned to the 16 Primary Factor scales and were subsequently suitable for item-level differential item functioning analysis. The remaining 30 items were not analysed due to uncertainty of item assignment to the correct scale.

Ordinal logistic regression was used to detect differential item functioning, using SPSS 15 PLUM. Ordinal logistic regression provided the probability of endorsing an item as a function of group members when matched for total amount of a latent trait. Equivalent amount of a latent trait will be assumed from identical scores. The first DIF analysis conducted was males (reference group) compared with females (focal group). The second DIF analysis conducted was New Zealand European (reference group) compared with Maori (focal group). Each reference group (Male, New Zealand European) had a dummy code 0, and each focal group (Female, Maori) had a dummy code of 1. The items identified as belonging to each of the 16 Primary Factor scales were summed to provide an individual's overall trait level for each individual scale. Items which display a – 2 log likelihood $p < 0.01$ will be identified as displaying some degree of DIF. An Estimate statistic $p < 0.01$ will provide the ordered log-odds (logit) regression coefficient. This odds ratio will provide the positive or negative probability direction of endorsement for the focal group (Female, Maori) in relation to the reference group (Male, New Zealand European), as either more likely or less likely to endorse an item in relation to the reference group when overall trait level is held constant.

3 Results

A t-test for independent means, significance $p < 0.01$ level was used to measure any potential effects based on gender or ethnicity at the overall 16 Primary Factor scale level. An ordinal logistic regression was run on each of the 16 Primary Factor scales, for each of the two group comparisons (male compared with female ; New Zealand European compared with Maori) which identified the number of items that displayed DIF. Items which displayed a – 2 log likelihood $p < 0.01$ were identified has displaying some degree of DIF. The number of items which achieved this level for each of the 16 Primary Factor scales were counted, and used to identify the percentage of items in each 16 Primary Factor scale displaying some degree of DIF. The same process was used with the Estimate statistic $p < 0.01$ to identify the percentage of items that were either more likely or less likely to be endorsed by the focal group in relations to the reference group when over all trait level was held constant.

Table 1 shows the number of items identified for each of the 16 Primary Factor scales for the current study out of a possible twelve. Table 1 shows the percentage of items identified (Appendix II) and the Cronbach's alpha measure of internal consistency of the 16 Primary Factor scales developed for this study. To provide context, Table 1 also provides three additional sources of internal consistency for the 16 Primary Factors scales reported in the 15FQ+ Technical Manual (2002) and Tyler and Newcombe (2006). Table 1 shows a significant range of internal consistencies developed for this study ranging .39 to .76, with an overall Cronbach's alpha of .78. The internal consistencies for the 16 Primary Factor scales range from unacceptable through to very good, whilst the overall internal consistency is very good.

Table 1

Cronbach's Alpha Internal Consistencies for the 15FQ+ 16 Primary Factor Scales

15FQ+ 16 Primary Factors Scales	No. Items Identified	Cronbach's Alpha Current Study (*n = 4798)*	15FQ+ Technical Manual (professional managerial *n* = 939)	Tyler & Newcombe (2006) Hong Kong Sample (*n* = 437)	15FQ+ Technical Manual UK Sample (*n = 325)*
Expedient	9 (75%)	0.76	0.82	0.72	0.81
Hard-headed	12 (100%)	0.58	0.74	0.64	0.77
Concrete	12 (100%)	0.64	0.72	0.64	0.79
Conventional	12 (100%)	0.48	0.76	0.61	0.79
Informal	12 (100%)	0.73	0.76	0.61	0.76
AffectedByFeelings	12 (100%)	0.76	0.78	0.76	0.77
Self-Assured	12 (100%)	0.60	0.82	0.73	0.83
Composed	12 (100%)	0.64	0.79	0.83	0.81
Distant Aloof	12 (100%)	0.73	0.74	0.72	0.78
Retiring	9 (75%)	0.39	0.82	0.83	0.81
Group-Oriented	12 (100%)	0.75	0.74	0.74	0.78
Low Intellectance	12 (100%)	0.42	0.77	0.80	0.80
Accommodating	6 (50%)	0.46	0.74	0.67	0.79
Direct	8 (66.7%)	0.63	0.76	0.61	0.78
Trusting	8 (66.7%)	0.56	0.75	0.68	0.77
SoberSerious	10 (83.3)	0.62	0.75	0.72	0.78
Total	170 (85%)	0.78	*not reported*	*not reported*	*not reported*

Table 2 shows how the different subgroups performed at the overall test level for each of the 16 Primary Factors scales. Table 2 shows the results of a t-test for independent means $p < 0.01$. It shows whether or not there are significant differences in the means between each sub groupings for each of the 16 Primary Factor scales.

Table 2

15FQ+ 16 Primary Factor Scales Means and Standard Deviations by Ethnicity and Gender

	Ethnicity						*Gender*					
	NZ European		*Maori*				*Male*		*Female*			
16 Primary Factors	M	SD	M	SD	*T*	*d*	M	SD	M	SD	*t*	*d*
Accommodation	10.94	1.90	10.71	2.00	-2.70*	0.12	11.07	1.90	10.66	1.90	-7.36*	0.22
Affected by Feelings	28.19	5.16	27.03	5.39	-5.03*	0.22	28.39	5.06	27.53	5.39	-5.60*	0.16
Composed	27.30	4.24	27.12	4.45	-0.91	0.04	27.96	4.12	26.22	4.28	-14.06*	0.41
Concrete	21.83	4.49	22.10	4.39	1.36	-0.06	22.14	4.45	21.43	4.50	-5.36	0.16
Conventional	24.81	3.58	23.42	3.61	-8.75*	0.39	24.34	3.57	25.11	3.63	7.19*	-0.21
Direct	15.80	3.50	15.35	3.41	-2.88*	0.13	15.54	3.38	16.06	3.64	5.05*	-0.15
Distant Aloof	17.67	4.05	17.30	3.73	-2.07*	0.10	18.51	4.21	16.25	3.23	-19.83*	0.60
Expedient	13.41	4.15	13.26	3.81	-0.83	0.04	13.67	4.23	12.97	3.88	-5.82*	0.17
Group Oriented	28.08	5.26	27.06	5.06	-4.41*	0.20	27.85	5.38	28.14	5.02	1.86	-0.06
Hard-headed	23.25	4.33	22.41	4.16	-4.42*	0.20	21.59	4.12	25.57	4.28	34.80*	-0.95
Informal	17.28	4.54	15.96	3.59	-6.74*	0.32	16.76	4.32	17.68	4.62	7.01*	-0.21
Low Intellectance	20.67	3.04	20.80	3.43	1.00	-0.04	20.59	3.04	20.82	3.17	-2.56*	-0.07
Retiring	19.28	2.91	18.88	3.00	-3.10*	0.14	19.11	2.93	19.43	2.91	3.78*	-0.11
Self-assured	19.39	4.09	19.02	3.90	-2.06*	0.09	19.20	3.94	19.58	4.24	3.22*	-0.09
SoberSerious	19.00	3.60	18.61	3.52	-2.50*	0.11	19.37	3.66	18.31	3.39	-10.11*	0.30
Trusting	19.29	3.09	17.44	3.50	-13.27*	0.56	18.96	3.19	19.23	3.22	2.86*	-0.08

*p < 0.01

Table 2 shows the means, standard deviations, significance tests, and effect sizes (Cohen's *d*; Cohen, 1988) for the 16 Primary Factors scales that make up the 15FQ+. Table 2 shows a break down by 16 Primary Factor scale and males compared with females and New Zealand European compared with Maori. Table 2 shows which scales have a statistically significant mean difference effect based on group membership at the $p < 0.01$ level. Cohen's *d* has been used to show the size of the observed effects (Cohen, 1992). The largest *d* ratio (-0.95) was based on gender for men (M = 21.49) and women (M = 25.57) comparison on the Hard Headed scale. The Trusting scale shows the largest difference based on ethnicity (*d* = 0.56) with Maori scoring lower (M = 17.44) than New Zealand European (M = 19.29). The majority of *d* ratio's for the 16

primary factors scales are less than $d = .50$ with the exception of Trusting (Ethnicity, $d = 0.56$), Hard Headed (Gender, $d = -0.95$) and Distant Aloof (Gender, $d = 0.60$). This suggests that despite the fact that a number of scales show statistically significant group mean differences the effects and differences between subgroups are small. The overall mean d ratios (0.25 for gender and 0.17 for ethnicity) indicate that gender differences are only marginally greater than ethnicity on the 16 primary factors scales.

The results of the DIF analysis based on gender are presented in Table 3, Table 4 and Table 5. The second column shows the number of items from each of the 16 Primary Factor scales that were identified and analysed as a percentage based on the total number of items each scale has. Of the 200 items that comprise the 16 primary factor scales, 170 items were identified (85%). An items was consider as displaying a degree of DIF, when the OLR was run and the – 2 Log Likelihood and Estimate statistic achieved $p < 0.01$.

Table 3

15FQ+ 16 Primary Factor Scales Number of Items Showing DIF Based on Gender

16 Primary Factors scale	Number of Items Identified	Total Number of Items displaying DIF Gender	Focal Group (female) More Likely to Endorse	Focal Group (female) Less Likely to Endorse
Accommodation	6 (50%)	(2)33.33%	(1)16.67%	(1)16.67
Affected by Feelings	12 (100%)	(8)66.67%	(5)41.67%	(3)25.00%
Composed	12 (100%)	(12)100%	(4)33.33%	(8)66.67%
Concrete	12 (100%)	(9)75.00%	(3)25.00%	(6)50.00%
Conventional	12 (100%)	(5)41.67%	(2)16.67%	(3)25.00%
Direct	8 (66.7%)	(5)62.50%	(2)25.00%	(3)37.50%
DistantAloof	12 (100%)	(9)75.00%	(6)50.00%	(3)25.00%
Expedient	9 (75%)	(3)33.33%	(2)22.22%	(1)11.11%
GroupOriented	12 (100%)	(8)66.67%	(3)25.00%	(5)41.67%
Hard Headed	12 (100%)	(11)91.67%	(8)66.67%	(3)25.00%
Informal	12 (100%)	(8)66.67%	(4)33.33%	(4)33.33%
Low Intellectance	12 (100%)	(4)33.33%	(2)16.67%	(2)16.67%
Retiring	9 (75%)	(4)44.44%	(2)22.22%	(2)22.22%
Self Assured	12 (100%)	(8)66.67%	(5)41.67%	(3)25.00%
SoberSerious	10 (83.3)	(7)70.00%	(2)20.00%	(5)50.00%
Trusting	8 (66.7%)	(3)37.5%	(2)25.00%	(1)12.50%
Total	170 (85%)	(106)62.35%	(53)31.18%	(53)31.18%

Table 3 shows the total number of items identified with DIF based on gender. Table 3 also shows which direction the DIF is for the focal group (female), either more likely to endorse or less likely to endorse an individual item in relation to the reference group (male) when overall trait level is held constant. Table 3 shows that the Composed/Tense Driven scale shows the most number of items identified with DIF with 100% (12) displaying some degree of DIF. It suggests that all items identified displayed some degree of DIF based on gender. The Composed/Tense Driven scale shows 33.33% (4) of items are more likely to be endorsed by females, and 66.67% (8) are less like to be endorsed by females, when matched for overall trait levels when compared with males. The scale Hardheaded/Tender Minded also shows a number of items which display

DIF based on gender as well. However in this instance, 66.67% (8) of items are more likely to be endorsed by the focal group (female). The scale Accommodation/Dominant shows the least amount of items with DIF, with 33.33% (2) of items displaying some degree of DIF. The Expedient/Conscientious and Trusting/Suspicious scales also show very little DIF. Of the 170 items identified 106 (62.35%) showed some degree of DIF based on gender. Table 3 also shows that the items which displayed DIF were evenly spread between the focal group (female) with 31.18% (53) being more likely or less likely to be endorsed by the focal group (female) when compared with the reference group (male) and being match on individual trait levels.

Table 4 shows the results of the DIF analysis based on ethnicity. It outlines the number of items identified for each of the 16 Primary Factor scales. Table 4 shows the total number of items identified as displaying some degree of DIF, and whether the focal group (Maori) are more likely or less likely to endorse the items which displayed DIF.

Table 4

15FQ+ 16 Primary Factor Scales Number of Items Showing DIF Based on Ethnicity

16 Primary Factors scale	Number of Items Identified	Total Number of Items Displaying DIF Ethnicity	Focal Group (Maori) More Likely to Endorse	Focal Group (Maori) Less Likely to Endorse
Accommodation	6 (50%)	(2)33.33%	(1)16.67%	(1)16.67
Affected by Feelings	12 (100%)	(4)33.33	(2)16.67%	(2)16.67
Composed	12 (100%)	(3)25%	(2)16.67%	(1)8.33%
Concrete	12 (100%)	(5)41.67%	(3)25.00%	(2)16.67
Conventional	12 (100%)	(6)50%	(3)25.00%	(3)25.00%
Direct	8 (66.7%)	(4)50%	(3)37.50%	(1)12.50%
Distant Aloof	12 (100%)	(2)16.67%	(1)8.33%	(1)8.33%
Expedient	9 (75%)	(3)33.33%	(1)16.67%	(2)22.22%
Group Oriented	12 (100%)	(4)33.33%	(2)16.67%	(2)16.67
Hard Headed	12 (100%)	(8)66.67%	(5)41.67%	(3)25.00%
Informal	12 (100%)	(4)33.33%	(2)16.67%	(2)16.67
Low Intellectance	12 (100%)	(10)83.33%	(4)33.33%	(6)50.00%
Retiring	9 (75%)	(1)11.11%	(1)11.11%	(0)00.00%
Self Assured	12 (100%)	(7)58.33%	(3)25.00%	(4)33.33%
Sober Serious	10 (83.3)	(3)30.00%	(1)10.00%	(2)20.00%
Trusting	8 (66.7%)	(3)37.50%	(2)25.00%	(1)12.50%
Total	170 (85%)	(69)40.59%	(36)21.18%	(33)19.41%

Table 4 shows the number of items identified with DIF based on ethnicity. Table 4 shows that Retiring/SociallyBold scale shows the least number of items with DIF, with 11.11% (1) displaying some degree of DIF. The Low Intellectance/High Intellectance scales shows the most number of items identified with differential item functioning, with 83.33% (8) of items showing DIF in relation to the reference group (New Zealand European) when matched on overall trait level. The majority of scales show very little DIF based on ethnicity, compared with DIF analysis based on gender. Overall 40.59% (69) of items display some degree of DIF based on ethnicity. Table 4 shows that based on ethnicity, the focal group (Maori) are more likely to

endorse 21.18%(36) of items, and less likely to endorse 19.41%(33) of items in relation to the reference group (New Zealand European) when matched on overall trait levels.

Table 5 shows the number of items that were identified for each of the 16 Primary Factor scales. Table 5 also shows the number of items that displayed no DIF across gender or ethnicity. This indicates the number of items that show no DIF. The number of items which displayed DIF across gender and ethnicity is also reported in Table 5.

Table 5

15FQ+ 16 Primary Factor Scales Number of Items that Show No DIF and Cross Group DIF

16 Primary Factors scale	Number of Items Identified	Number of items that show no DIF across Gender or Ethnicity	Cross Group DIF items
Accommodation	6 (50%)	(2) 33.33%	0.0%(0)
Affected by Feelings	12 (100%)	(2)16.67%	17.00%(2)
Composed	12 (100%)	(0) 00.00%	25.00%(3)
Concrete	12 (100%)	(1)8.33%	25.00%(3)
Conventional	12 (100%)	(3)25.00%	17.00%(2)
Direct	8 (66.7%)	(2)25.00%	38.00%(3)
Distant Aloof	12 (100%)	(2)16.67%	8.00%(1)
Expedient	9 (75%)	(4)44.44%	11.00%(1)
Group Oriented	12 (100%)	(2)16.67%	17.00%(2)
Hard Headed	12 (100%)	(0) 00.00%	58.00%(7)
Informal	12 (100%)	(3)25.00%	25.00%(3)
LowIntellectance	12 (100%)	(1)8.33%	33.00%(4)
Retiring	9 (75%)	(4)44.44%	0.0%(0)
Self Assured	12 (100%)	(2)20.00%	42.00%(5)
SoberSerious	10 (83.3)	(2)20.00%	20.00%(2)
Trusting	8 (66.7%)	(4)50.00%	25.00%(2)
Total	170 (85%)	(34)20.00%	23.53%(40)

Table 5 shows that of the 170 items identified in this study 20.00% (34) of items display no DIF. The scale Expedient/Conscientious (44.44%), Retiring/SociallyBold (44.44%), and Trusting/Suspicious (50.00%) show the most number of items without DIF. The scales Composed/TenseDriven (00.00%), HardHeaded/TenderMinded (00.00%) and LowIntellectance/HightIntellectance (8.33%) scales have the least number of items that do not display DIF. Table 5 also shows that there is very little cross loading of DIF, with 23.53% (40) items showing DIF for both gender and ethnicity. The scale HardHeaded/TenderMinded (58.00%) and SelfAssured/Apprehensive (42.00%) shows the most number of items with large across group DIF. The scales Retiring/SociallyBold (00.00%) and Accommodation/Dominant (00.00%) show the least amount of cross-group DIF.

4 Discussion

4.1 Introduction

Personality testing as part of pre-employment selection is continuing to increase worldwide (Sheppard et al. 2006; Jenkins, 2001). However, the examination of measurement bias in pre-employment personality inventories has lagged behind that of educational and psychopathological measurement tools (Sheppard et al. 2006; Roznowski & Reith, 1999; Sackett & Wilk, 1994). The purpose of this study was to add to the understanding of measurement bias in pre-employment selection tools. This study has evaluated potential item level bias based on gender and ethnicity subgroups in a pre-employment personality test using differential item functioning (DIF). This study represents an important step in the effort to evaluate DIF by gender and ethnicity. More importantly this study has used a large New Zealand sample of male and females who identify themselves as New Zealand European or Maori.

A number of different analyses have been conducted in this study to analyse the 15FQ+ (Psytech, 2002) for gender and ethnicity differences. Analysis at the mean subgroup level comparing each of the 16 Primary Factors scales suggest that subgroup differences exist, but are small. The scale Trusting/Suspicious showed the biggest group mean difference (d = 0.56) for ethnicity, whilst the HardHeaded/TenderMinded scale showed the biggest gender difference (d = -0.96). Using Cohen's d (Cohen, 1988) to measure the effect of mean differences in the 16 Primary Factor scales between males and females, and New Zealand European and Maori suggest effects remain minor.

The results of DIF analysis show that a number of items contained in the 15FQ+ display some degree of DIF based on gender or ethnicity. Differential item functioning analysis has identified that 62.35% (106) of items display some degree of DIF based on gender. Differential item functioning analysis based on ethnicity shows that 40.59% (69) of items show some degree of DIF. Overall 20.00% (34) of items displayed no DIF across gender and ethnicity. The Trusting/Suspicious scale showed the least amount of DIF overall. Interestingly, very little cross loading of DIF occurred across gender and ethnicity. That is to say, very few items displayed DIF both for gender and ethnicity. The scale which showed the most cross loading was Composed/TenseDriven (25.00%). The implications for this lack of cross loading are discussed later (Sheppard et al. 2006).

The scale Composed/TenseDriven when analysed for gender DIF displayed the highest number of items displaying DIF (100.00%). Based on gender females were less likely to endorse 66.67% (8) of items and more likely to endorse 33.33% of items compared with the reference group male. Of the 170 items identified 62.35% (106) items displayed some degree of DIF based on gender. Differential item functioning based on ethnicity showed a lower incidence of DIF, with 40.59% (69) of items displaying some degree of DIF. The scale LowIntellectance/HighIntellectance showed the most number of items identified with DIF based on ethnicity. Overall 83.33% (8) showed some degree of DIF, with 50.00% (6) of items being less likely to be endorsed by the focal group (Maori) when overall trait level was held constant. Overall, 40.59% (69) of items showed some degree of DIF, which taken across 16 scales suggests less significant DIF, and much less DIF compared with gender.

4.2 Implications

The results found in this study are in line with studies of a similar nature which have investigated DIF in personality measures. A number of studies have revealed that DIF is "...rampant..." (Kulas, Merriam, & Onama, 2008 p. 1103) across a number of psychometric tests and environments. Sheppard, Han, Colarelli, Dai and Kin (2006) found that over one third of the Hogan Personality Inventory (Hogan & Hogan, 1992) items display some degree of DIF across race and gender. Young and Sudweeks (2005) found that more than 40% of items in the Multidimensional Self Concept Scale (Bracken, 1992) items displayed some degree of DIF based on gender. Huang, Church, and Katigbak (1997) found that almost 40% of NEO-PI (Costa & McCrae, 1992) items displayed DIF based on nationality. Waller, Thompson, and Wenk (2000) found that close to 40% of items in the Minnesota Multiphasic Personality Inventory (Butcher, Dahlstrom, Graham, Tellengen, & Kaemmer, 1989) displayed DIF based on race. This study has found that 40.59% of items display DIF based on ethnicity, whilst 62.35% of items displayed DIF based on gender. These results suggest that the same issues that cause concern in other tests are also present in the 15FQ+.

Very few items display across-group DIF, for example this study found only 23.53% (40) of items loaded on both gender and ethnicity. This raises the same concerns Sheppard et al. (2006) have, that moderately biased items may display some bias for one category, but not another. For example, an item may show bias for gender subgroups, but not on different ethnic subgroups. This would seem supported by the current study which found only a small number of items show bias for both gender and ethnicity. Eliminating an item which displays bias based on gender does

not guarantee that the test will be free of item bias by ethnicity. The opposite is also true, an item may be analysed for gender DIF, and display none, but that is not to say that it will not show DIF based on ethnicity. For example, this study found that for the scale Composed/TenseDriven 100% (12) of items displayed some degree of DIF based on gender, however based on ethnicity only 25% (3) items displayed DIF. The reverse is true of the scale LowIntellectance/HighIntellectance, with 83.33% (10) items displaying DIF for ethnicity, and only 33.33% (4) of items displaying DIF based on gender.

By analysing across-group DIF matters are further complicated. For example the Accommodation/Dominant and Retiring/SociallyBold scales show no cross group DIF. However, when reviewed based on the gender and ethnicity subgroups Accommodation/Dominant and Retiring/SociallyBold each show 33.33% (2) of items display DIF. However, because there is no cross DIF, this means that the same two questions are not biased for the same subgroup in these scales. Differential item functioning analysis such as this study, and Sheppard et al. (2006) have found very little incidence of DIF cross loading, when two different sub-grouping analyses have been conducted on the same test. The many complexities of subgroup membership and the nearly *ad infinitum* combinations of group membership suggest that it would be extremely difficult to achieve a sufficient item pool which displayed no bias in one form or another.

There are a countless number of possible groups that an item may still be biases toward. It simply has not been compared or subject to a DIF analysis. Additionally individuals can be sub-grouped in to multiple classifications such as gender, ethnicity, language, and culture which

themselves have numerous classifications. For example this study used Maori Male participants, however, participants were not analysed as a Maori Male participants. Rather individuals were analysed based on their ethnicity and gender separately. So an item which displayed gender DIF in this study, when compared with Maori men and Maori women may potentially display no DIF. The reverse may also be true, an item may not have shown DIF based on gender, however, by adding an additional layer of classification such as Maori and Male the item may show DIF when compared with Maori and female.

Assuming that the results of this study and of Sheppard et al. (2006), Waller et al. (2000), and Young and Sudweeks (2005) are representative, this raises several issues, and questions for consideration. It raises questions about the use and development of personality inventories, for example what should be done with biased items. One may ask, is measurement bias a serious problem for personality inventories, what is the utility of DIF and is it a worth while process. How can differences be controlled for between subgroups, if at all? Some research suggests that DIF is not serious in a practice, and see it as a source of additional information and insight in to subgroup behaviour, whilst others argue items should be removed, and scales reanalysed and further developed.

So with the purported prevalence of DIF existing in a variety of tests in a variety of different psychological domains, what is one to make of it all? One solution is to view the biased items as a serious problem, and to continue analysing tests for DIF, and removing items from that test, and then reanalysing the test, or including new items which do not display DIF. In commercially well developed and marketed, used and distributed tests this process is highly problematic. Tyler

and Newcombe (2006) suggest the need to asses problematic items through item analysis and either delete or refine them. However, removal of items can cause problems with established tests. This is especially difficult in tests which are computer-scored or internet-based such as the 15FQ+. Potential consequences of DIF include some form or contribution to adverse impact, by lowering certain groups mean scores due to the relationship bias that exist with items that display DIF. Additionally the generalisability of a test is reduced across sub group populations (Schmit, Kihm, Robie, 2000). Despite the purported prevalence of DIF the resulting impact on related selection and assessment decisions may not be significant (Kulas, Merriam, & Onama, 2008; Stark, Chernyshenko & Drasgow 2004). Roznowski and Reith (1999) and Waller et al. (2000) suggest that DIF does not necessarily translate in to differential test functioning. Therefore the removal of biased items may have little practical impact in improved test performance and quality.

Roznowski and Reith (1999) reported that the measurement quality of tests was not significantly or seriously diminished when items which displayed DIF were retained. This hypothesis, though on the face of it illogical has been supported by Reise, Smith, and Furr (2001) who identified gender DIF in the NEO PI-R Neuroticism Scale and Waller et al. (2001) who indentified race based DIF in the MMPI. Both suggest that the overall effect at the scale level remained inconsequential. This study would appear to support this claim, and are in line with the findings of Sheppard et al. (2006) and Ones and Anderson (2002) who found that overall group mean differences between gender and ethnicity on scales despite being statistically significant are still small. These studies represent an important step forward in themselves, as Roznowski and Reith

(1999) report that test bias and item bias are studied and discussed as separate phenomena, when really them remain closely interrelated.

This study has shown there are a number of ways to detect differential item functioning, the Generalised Mantel-Haenszel procedure, IRT, logistic regression, SIBTEST and a host of other statistical techniques. Whilst there is an abundance of cost and time effective means for detecting DIF, as of yet no uniformly accepted or adequate theory exists to explain or predict which items will display DIF or not, and what is causing it. Given that a number of studies suggesting that a number of tests display some degree of DIF across three domains of psychological testing (Educational, Abnormal, and Industrial and Organisational) it would suggest that items which display a moderate degree of DIF do not appear to influence test quality or utility and validity. Therefore the removal of items exhibiting some degree of DIF would yield little advantage in most circumstances (Reise, Smith, & Furr, 2001; Waller et al. 2001; Zumbo, 1999).

So what is the point of all this DIF if at the overall test level there appears to be no differential test functioning and there are so many subgroup combinations that a pool of total unbiased items seems impossible? One area of importance is that DIF analysis contributes significantly to the development, content, and construction of tests and is increasingly becoming an integral part of test production and evaluation (Holland & Wainer, 1993). Differential item functioning remains one of many tools, which inform test development and validity. It seems likely that items which display an extreme form of DIF when removed increase test performance with more validity and less bias (Sheppard et al. 2006). Additionally, measurement irrelevancies may give further

insight in to subgroup behaviour, and be a source of additional investigation and study in to what is causing it. It is important to reiterate that if an item is biased or extremely offensive to both subgroups under consideration then DIF will not detect this, as it only detects differential responses. An item which is grossly biased or offensive to both groups, will show a similar response pattern, and is therefore better detected though subject matter experts and content analysis (Sheppard et al 2006). Further to this, DIF alone is not enough to reject an item, or a test (Zumbo, 1999).

4.3 Limitations

One of the strengths of this study is the population of interest, male and females, who identify as either New Zealand European and Maori. To date DIF studies have not been conducted using these participants or subgroup combinations. This also limits the study as well, as the results may not generalise across other groups. As mentioned earlier there a many possible subgroup combinations and the low rate of cross item DIF found in this study and others (Sheppard et al. 2006) suggests that each subgroup comparison has an element of uniqueness about it, which makes comparisons to other groups difficult. For example New Zealand male and female participants DIF items on this test may not generalise to another set of male and female participants in China who have completed their language version of the test (Tyler and Newcombe, 2006). Additionally, despite the support for the universality of a Five Factor Model, idiosyncrasies among different ethnic subgroups will exert some degree of influence when conducting DIF. This leads to the problem mentioned earlier, of the impracticality of completing numerous group combinations of DIF. The results of this study can be generalised to the New

Zealand population and the use of the 15FQ+, however, it may not be suitable to generalise these results across other populations, such as Caucasian Americans and African Americans or Chinese participants.

This study used a sample of New Zealand European and Maori participants. A significant issue in the current study is the number of participants who identified as Maori compared with the number of participants who identified themselves as New Zealand European. Of the 4798 participants 87.9% were New Zealand European compared with 12.1% who identified themselves as Maori. Whilst these figures are loosely representative of the population in New Zealand, the large number of New Zealand European participants is liable to influence the statistical power in the analysis conducted. The large number of New Zealand European participants is likely to exert some influence over the results of this study and magnify any true difference. Additionally such a large participant pool is likely to magnify any real differences that do exist, and make significant results easier to achieve. A further limitation of this study is the assignment of items to the 16 Primary Factor scales. It is feasible that items may have been assigned to the incorrect scales. Some scales also have lower item identification rates, as low as 50% in the case of Accommodation/Dominant. This may influence the magnitude of differential item functioning. Additionally it is unlikely that all 16 Primary Factors scales were unidimensional. This is liable to influence and distort the results as well.

Further to this, the 15FQ+ scales developed for this study show a number of scales with reliability less than the conventional .70 suggesting that some items may have been incorrectly assigned to their respective scales. Taken within the context of the 15FQ+ technical manual and

Tyler and Newcombe (2006), it would seem to suggest that the scales of the 15FQ+ are difficult to achieve extremely reliable internal consistency results. Overall the scale developed for this study achieved an internal reliability of .76 suggesting it had an overall degree of internal reliability whilst some individual 16 Primary factor scales may have lack internal consistency. Additionally Tyler and Newcombe (2006) suggest that the 15FQ+ requires further development work to enhance the reliability of the 16 Primary Factor scales. A final cautionary note to this study is not to interpret the rate of differential item functioning at the overall level. Differential item functioning analysis was only conducted on 85% of items (170 out of 200) across 16 different scales.

4.4 Future research

An area of improvement and further research includes an investigation of what is causing DIF. Some form of factor analysis or content analysis which investigates themes among items which display DIF may shed additional light on the 15FQ+ and its behaviour at the subgroup level. Further to this the magnitude and effect size of DIF contained in the 15FQ+ could be investigated to see how the removal of certain items might influence the performance of the test. Though DIF items have been identified, how participants are possibly advantaged or disadvantaged or why some participants respond differently remains largely unexplored, in this study and in general (Sheppard et al. 2006). By focusing on the unrelated variance a greater understanding of what may be causing DIF could be gained (Webb, Cohen, & Schwanenflugel, 2008).

As has been clearly stated, there are a host of different group combinations that could be completed. Areas that would warrant further investigation include additional studies which are carried out to determine if similar levels of DIF are found on the 15FQ+ using different populations of interest. How New Zealand European and Maori compare with a host of other different ethnic groups could be completed as well. Additionally a study could be completed which investigates different language versions of the 15FQ+. For example participants who have completed the Traditional Chinese version of the 15FQ+ could be compared with participants who have completed the 15FQ+ English version (Tyler and Newcombe, 2006). As the area of personality inventories and DIF has been limited to date, further research should be conducted to see how other personality inventories such as the OPQ (Bartram et al., 2006) and MBTI (Myers, McCaulley, Quenk, Hammer, 1998) perform on DIF (Sheppard et al. 2006). The number of personality tests that exist, coupled with the number of group combinations, and lack of theory to explain why items display DIF suggests there is still a lot of work to be done in investigating DIF.

This study has made an important step forward in investigating potential measurement bias in personality inventories using a New Zealand sample. With the increasing use of personality in employee selection, it is important to ensure that any test used is investigated thoroughly to ensure its robustness (Sheppard et al. 2006; Zumbo, 1999). It is important to stress that personality measurement should never form the sole basis on which to make an employment decision (Sheppard et al. 2006; Roznowski & Reith, 1999). Personality inventories appear to exert a powerful influence in the world of employee selection at the present time (Jenkins, 2001). Major research to date has stress that whilst personality factors are important in job performance,

they should by no means form the sole decision making tool (Sheppard et al. 2006). Rather, psychometric testing gives a valuable insight in to an individual, and forms a basis for exploration and investigation. Personality testing in pre-employment selection should form but one key in identifying an individual who can add worth and value to the organisation.

5 References

Abedi, J., Leon, S., & Kao, J. (2007). *Examining differential item functioning in reading assessments for students with disabilities*. Minneapolis, MN: University of Minnesota.

Barrick, M R., & Mount, M. K. (1991). The Big Five personality dimensions of job performance: a meta-analysis. *Personnel Psychology*, 44(1), 1-26.

Bartram, D., Brown, A., Fleck, S., Inceoglu, I., & Ward, K. (2006). *OPQ32 Technical Manual*. Thames, UK: SHL Group plc.

Bracken, B. A. (1992). *Multidimensional Self Concept Scale Manual*. Austin, TX: Pro Ed.

Borsboom, D., Mellenbergh, G. J., & van Heerden, Jaap. (2002). Different kinds of DIF: a distinction between absolute and relative forms of measurement invariance and bias. *Applied Psychological Measurement*, 26, 433450.

Butcher, J. N., Dahlstrom, W. G., Graham, J. R., Tellegen, A., & Kaemmer, B. (1989). *The Minnesota Multiphasic Personality Inventory-2 (MMPI-2): Manual for administration and scoring*. Minneapolis, MN: University of Minnesota Press.

Cattell, R. B., (1946). *The Description and Measurement of Personality*. Yonkers-on-Hudson, NY: World Book Co.

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items an NCME instructional module. *Educational Measurement: Issues and Practice*, 17, 31-44.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112(1), 155-159.

Cohen, J. (1988). *Statistical power analysis for the behavioural sciences*. Hillsdale, NJ: Lawrence Erlbaum.

Collins, J. M., & Gleaves, D. H. (1998). Race, job applicants, and the Five-Factor Model of personality: implications for Black Psychology, Industrial and Organisational Psychology, and the Five-Factor Theory. *Journal of Applied Psychology*, 83(4), 531-544.

Collins, W. C., Raju, N. S., & Edwards, J. E., (2000). Assessing differential functioning in a satisfaction scale. *Journal of Applied Psychology*, 85(3), 451-461.

Condon, L., Morales-Vives, F., Ferrando, P. J., & Vigil-Colet, A., (2006). Sex differences in the Full and Reduced versions of the Aggression Questionnaire: a question of differential item functioning? *European Journal of Psychological Assessment*, 22(2), 92-97.

Costa, P. T., & McCrae, R. R. (1992). *NEO PI-R professional manual*. Odessa, FL: Psychological Assessment Resources, Inc.

Day, D. V., & Silverman, S. B. (1989). Personality and job performance: evidence of incremental validity. *Personnel Psychology*, 42, 25-36.

Englert, P. (2006). May 9th 2006 Industrial and Organisation Special Interest Group: The Myth of Psychometrics. Retrieved August 29, 2008, from http://www.hirinz.org.nz

Escorial, S., & Navas, M. J. (2007). Analysis of the gender variable in the Eysenck Personality Questionnaire Revised scales using differential item functioning techniques. *Educational and Psychological Measurement*, 67, 990-1001.

Fidalgo, A. M., Hashimoto, K., Bartram, D., & Muniz, J. (2007). Empirical Bayes versus standard Mantel-Haenszel statistics for detecting differential item functioning under small sample conditions. *The Journal of Experimental Education*, 75(4), 293-314.

French, B. F., & Maller, S. J. (2007). Iterative purification and effect size use with logistic regression for differential item functioning detection. *Educational and Psychological Measurement*, 67(3), 373-393.

Goldberg, L. R. (1990). An alternative "description of personality"; the Big Five factor structure. *Journal of Personality and Social Psychology*, 59, 1216-1229.

Hogan, R., & Hogan, F. (1992). *Hogan Personality Inventory manual*. Tulsa, OK: Hogan Assessment Systems.

Holland, P. W., & Wainer, H. (1993). *Differential Item Functioning*. Hillsdale, NJ: Lawrence Earlbaum Associates.

Holmes-Finch, W. & French, B. F. (2008). Anomalous type I error rates for identifying one type of differential item functioning in the presence of the other. *Educational and Psychological Measurement* 68(5), 742-759.

Huang, C. D., Church, A. T., & Katigbak, M. S. (1997). Identifying cultural differences in items and traits: differential item functioning in the NEO personality inventory. *Journal of Cross-Cultural Psychology*, 28, 192-218.

Hubley, A. M., & Zumbo, B. D. (1996). A dialect on validity: where we have been and where are we going. *The Journal of General Psychology*, 123(3), 207-215.

Jenkins, A. (2001). *Companies' use of psychometric testing and the changing demand for skills: a review of the literature*. London, UK: Centre for the Economics of Education, London School of Economics and Political Sciences

Kulas, J. T., Merriam, J., Onama, Y. (2008). Item-trait association, scale multidimensionality, and differential item functioning identification in personality assessment. *Journal of Research in Personality*, 42, 1102-1108.

Lord, F. (1980). *Applications of item response theory to practical test problems*. Hillsdale, NJ: Erlbaum.

McCrae, R. R., & Costa, P. T. (1997). Personality trait structure as a human universal. *American Psychologist*, 52(5), 509-516.

McCrae, R. R., & Costa, P. T. (1985). *Updating Norman's "Adequate Taxonomy"; intelligence and personality dimensions in natural language and in questionnaires*. Journal of Personality and Social Psychology, 43(3), 710-721.

Mitchell, J. (1999). *Measurement in Psychology*. Cambridge, UK: Cambridge University Press.

Myers, I. B., McCaulley M. H., Quenk, N. L., Hammer, A. L. (1998). *MBTI Manual (A guide to the development and use of the Myers Briggs type indicator)*. Lexington, KY: Consulting Psychologists Press

Norman, W. T. (1963). Toward an adequate taxonomy of personality attributes: replicated factor structure in peer nomination personality ratings. *Journal of Abnormal and Social Psychology*, 66, 574-583.

NZCER, (2008). *Human Resource Assessments Catalogue 2008-2009*. Wellington, NZ: New Zealand Council for Educational Research.

Ones, D. S., & Anderson, N., (2002). Gender and ethnic group differences on personality scales in selection: some British data. *Journal of Occupational and Organisational Psychology*, 75, 255-276.

Psychometrics Limited (2002). *The 15FQ+ Technical Manual*. Pulloxhill, Bedfordshire, UK: Psychometrics Limited.

Roznowski, M., & Reith, J., (1999). Examining measurement quality of tests containing differentially functioning items: do biased items result in poor measurement? *Educational and Psychological Measurement*, 59, 248-269.

Reise, S. P., Smith, L., Furr, R. M. (2001). Invariance on the NEO PI-R neuroticism scale. *Multivariate Behavioural Research*, 36, 83-110.

Saad, S. & Sackett, P. R. (2002). Investigating differential prediction by gender in employment-oriented personality measures. *Journal of Applied Psychology*, 87(4), 667-674.

Sackett, P. R., & Wilk, S. L., (1994). Within-group norming and other forms of score adjustment in pre-employment testing. *American Psychologist*, 49, 929-954.

Scherbaum, C. A., & Goldstein, H. W. (2008). Examining the difference between race-based differential item functioning and item difficulty. *Educational and Psychological Measurement*,

68(4), 537-553.

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: practical and theoretical implications of 85 years of research findings. *Psychological Bulletin*, 124(2), 262-274.

Sheppard, R., Han, K., Colarelli, S. M., Dai, G. & King, D. W., (2006). Differential item functioning by sex and race in the Hogan Personality Inventory. *Assessment*, 13, 442-453.

Smith, L. L., & Reise, S. P. (1998). Gender differences on negative affectivity: an IRT study of differential item functioning on the Multidimensional Personality Questionnaire Stress Reaction Scale. *Journal of Personality and Social Psychology*, 75(5), 1350-1362.

Stark, S., Chernyshenko, O. S., Drasgow, F. (2004). Examining the effects of differential item functioning (functioning and differential) test functioning on selection decisions: when are statically significant effects practically important? *Journal of Applied Psychology*, 89, 497-508.

Swaminathan, H. & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, 27(4), 361-370.

Tett, R. P., Jackson, D. N., & Rothstein, M. (1991). Personality measures as predictors of job performance: a meta-analytic review. *Personnel Psychology*, 44(4), 703-751.

Thissen, D., Steinberg, L., & Gerrard, M. (1986). Beyond group-mean differences: the concept of item bias. *Psychological Bulletin*, 99(1), 118-128.

Timmerman, T. A. (2004). Relationships between NEO PI-R personality measures and job performance ratings of inbound call centre employees. *Applied Human Resource Management Research*, 9(1), 35-38.

Tyler, G. P., & Newcombe, P. A. (2006). Relationship between work performance and personality traits in Hong Kong organisational setting. *International Journal of Selection and Assessment*, 14(1), 37-50.

Waller, N. G., Thompson, J. S., & Wenk, E. (2000). Using IRT to separate measurement bias from true group differences on homogeneous and heterogeneous scales: an illustration with the MMPIS. *Psychological Methods*, 5, 125-146.

Webb, M. L., Cohen, A. S., & Schwanenflugel, P J. (2008). Latent class analysis of differential item functioning on the Peabody Picture Vocabulary Test III. *Educational and Psychological Measurement*, 68(2), 335-351.

Young, E. L., & Sudweeks, R. R. (2005). Gender differential item functioning in the Multidimensional Self Concept Scale with a sample of early adolescent students. *Measurement and Evaluation in Counselling and Development*, 38, 29-44.

Zumbo, B. D. (1999). *A Handbook on the Theory and Methods of Differential Item Functioning (DIF): Logistic Regression Modelling as a Unitary Framework for Binary and Likert-Type (Ordinal) Item Scores.* Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defence.

5 Appendices

Appendix A:

CANDIDATE INFORMATION AND RELEASE FORM

Selection

Please take a few minutes to read through the following forms once they have been explained to you, to get an understanding of the assessment process.

- The completion of assessment exercises is usually required to assist decision-making in areas of selection and recruitment, managerial development, career path planning, and training and development.

- The reporting of information to the candidate is in verbal summary form and is specific to the position under consideration. Results are considered valid for around 12 months, after which time data are made anonymous.

- Often, assessment tools used for selection revolve around the measurement of your personality (i.e. interpersonal style, work attitudes, career drivers) and problem solving ability. Your performance on any test may then be compared against several normative groups or an ideal profile that was developed for the particular job.

- Assessment tools are used in conjunction with a range of other selection methods such as the interview, CV, and other application forms to help make accurate, reliable, and robust selection decisions.

- You may be required to spend between 0.5 to 4 hours completing the assessment programme; your consultant will advise you of the expected time. This session will result in the writing up of a report summarising your performance. This will then be used in helping make a final selection decision, used in conjunction with the other tools mentioned above.

Please read through the following information and sign if you agree to the conditions of the assessment:

- I agree that the results of the evaluation completed today will be released to:

_____

(Name of company where role is being applied for)

Along with other sources of data, the assessment testing process is to assist in making a selection decision. The results of an assessment are given to the above named recipient to use as a guide only (in the recruitment process) and are not recommended as the sole factor on which selection decisions are made.

(Please turn over to the next page)

- I understand that **** (the test supplier) will receive the database containing my results, and that it will be stored in a collective database for statistical use only, with my name held strictly confidential. Published statistics will not be traceable back to me.

- In line with the Privacy Act of 1993, I understand that because of the evaluative nature of the material obtained, I may not have full access to the results in detail. However, I understand that I will be entitled to verbal feedback as part of the assessment process (in line with the New Zealand Psychological Society Ethics, clause 2.1.4). This will be offered either from a consultant at ****, or by the client who has commissioned this assessment session.

- In line with section 29 of the Privacy Act of 1993, I understand that if as part of my assessment I have been asked to complete the Stanton Survey of Integrity, no feedback is available on my results. This is on the advice of the test publishers, **** is obliged to follow these recommendations.

- The information I provide will be secured against loss, unauthorised access, modification, disclosure, and misuse.

- I understand that personality assessments include a measure of the extent to which the answers given are a true representation of myself.

☐ I have told the test administrator of any physical, health or other issue that may impact on my performance
☐ I have been given the option to complete the assessments either by paper and pencil or PC
☐ I have enough lighting in the testing room to complete the evaluation
☐ I have reading glasses/contact lenses if required
☐ I am aware of the approximate length of time this evaluation will take
☐ I am aware of the nature of the evaluations that I will be undertaking
☐ I am aware that I am not entitled to use a calculator or dictionary in the assessments unless instructed by the administrator
☐ I have switched off my mobile phone (if I have one with me)

If there is any reason why today is not an appropriate time for me to be undertaking the assessment (e.g. due to personal illness or stress), then I have let the test administrator know (or will do so as soon as possible).

I have read, understand, and agree with the information and requirements relating to my assessment session.

Name:_____

Date:_____

Signature:_____

Contact: (Phone Number/Email Address)_____

Appendix B:

Assignment

The 15FQ+ is a commercially available test, therefore requests for a scoring key and knowledge of which items constitute each scale was declined. Therefore before conducting any analysis of the items, they were assigned to their respective 16 Primary Factors scales.

Items were assigned to their respective 15FQ+ 16 Primary Factor scales using (a) information contained in the 15FQ+ Technical Manual (Psytech, 2002), (b) intuitive judgement based on face validity, (c) factor analysis, and (d) item scale statistics. The 15 FQ+ Technical Manual (Psytech, 2002), indicated a number of items and their respective scales. For example on the scale, Low Intellectance/High Intellectance, individuals with Low Intellectance are describe as individuals who "...find is confusing when people use long words." (15FQ+ Technical Manual, Psytech, 2002 p.14), therefore the item "I dislike it when people use complicated words", was taken at face value as loading on the Low Intellectance/High Intellectance scale.

The 15FQ+ is reported as mapping on to five Global Factors (Extraversion/Introversion, Low Anxiety/High Anxiety, Pragmatism/Openness, Independence/Agreeableness, Low Self-Control/High Self Control), similar to 'The Big 5', but in no way reported to have 'The Big 5' as its underlying structure. To establish if the five Global Factors wore borne out in this study, a factor analysis was run in line with Stage 6 of the development of the 15FQ+ as reported in the 15FQ+ Technical Manual (Psytech, 2002). A factor analysis was run with the data set of 4798 participants from this study with all 200 items. The factor analysis methodology is not reported in the 15FQ+ Technical Manual (Psytech, 2002), therefore items were analysed using principal

components analysis with varimax rotation. Principle components analysis with varimax rotation was used in an attempt to reduce the amount of intercorrelation experienced when running the factor analysis. A five factor solution supported the underlying Five Global Factors scales. Component loadings of items served as an additional guide to assign items to their respective scales on the 16 Primary Factor scales. For the purpose of comparison additional factor analysis were run using principle components analysis with direct oblumin rotation, and principle axis factoring with direct oblumin rotation, which allow for intercorrelation. Each revealed very similar patterns of item clusters which matched the initial factor analysis solution. This would suggest that the initial orthogonal rotation was acceptable in light of the later run oblique rotations which also displayed low correlations.

The factor analysis supported the extraction of the 5 Global Factors (Extraversion/Introversion, Low Anxiety/High Anxiety, Pragmatism/Openness, Independence/Agreeableness, Low self-control/High self-control). Though closely related to the 'Big 5', the 15FQ+ does not map perfectly on to this model, nor does it claim to. A factor analysis was also conducted in an attempt to achieve a 16 factor solution (principle components analysis with varimax rotation), which was unable to extract the 16 Primary Factors scales. During the construction of the 15FQ+ as reported in the 15FQ+ Technical Manual (Psytech, 2002), the items comprising the 16 Primary Factor scales were not subject to a factor analysis to achieve a 16 factor solution. Nor should this be expected, as due to issues of unidimensionality, and multidimensionality it would be highly unlikely that factor analysis would extract a 16 factor solution that was perfectly unidimensional on each scale. Rather the 15FQ+ Technical Manual (Psytech, 2002) reports that items comprising the 16 Primary Factor scales were created, and then scale statistics were

developed to measure their validity. A factor analysis was subsequently run (Step 6; 15FQ+ Technical Manual, 2002) on all items comprising the 16 Primary Factor scales to achieve an understanding of the underlying pattern and structure. This achieved the super-ordinate 5 Global Factors, similar to 'The Big Five' (Tupes & Christal, 1962).

A total of 170 items (85.00%) were assigned to their respective 16 Primary Factor scales, and were suitable for differential item functioning analysis. Thirty items (25.00%) could not be analysed for differential item functioning due to uncertainly of scale assignment. Despite some scales having fewer items than the original 16 Primary Factor scales, the internal consistency Cronbach's alpha of the newly derived scales for this study are comparable to those reported in Tyler and Newcombe (2006) and the 15FQ+ Technical Manual (Psytech, 2002). Explanations for newly derived scales which have lower internal consistency include (a) failing to identify critical items for a scale, (b) the incorrect assignment of items to scales, and (c) a lack of internal consistency among the identified items that comprise the newly derived scales for this sample.